LINGUISTIC RESEARCH AT THE RAND CORPORATION

David G. Hays[1]
The RAND Corporation

Summary

This paper describes postediting rules for description of funclion in context, work on computational routines for semi-automatic analysis, the concept of idiom-in-structure, and two broad problems on which work is just beginning at RAND: grammatic transformation and distributional semantics. The latter problems are especially important for automatic indexing, abstracting, and text searching.

Introduction

We are spending this winter in writing a major report. After nearly three years of research, and after processing a quarter-million running words of text, we find that we have a lot to say. In the self-description that we furnished the National Science Foundation for its most recent survey of MT studies,[2] we expressed our hope that we would have a completed system in operation during the summer of 1959. Our hope was fulfilled, in a manner of speaking, and we set out to describe what we had accomplished. As we write, however, we find that clear exposition highlights our every weakness. Our writing is therefore interlarded with efforts to eliminate some of the flaws. We are not the first to discover that writing a book takes longer than a reasonable man would dare guess, nor will we be the last.

Our table of contents has been revised several times. At first it was a sketch of what Russian-English MT requires: (1) a method of

[1] Other members of the RAND project include Kenneth E. Harper and Dean S. Worth (both of UCLA, consultants); Andrew S. Kozak, Dolores V. Mohr, Joan H. Pustula, and Barbara J. Scott (linguistic technicians); Theodore W. Ziehe, Hugh S. Kelly, and Charles H. Smith (programmers). The plural pronoun in the text includes these persons, but errors should be attributed to the author alone.

[2] Madeline M. Henderson, and Nancy Ripple, editors, Current Research in Scientific Documentation, No. 5 (October, 1959), Washington, D. C. : National Science Foundation, Office of Science Information Service, NSF-59-24. (Cf. p. 60.)

semi-automatic lexicographic and grammatic research; (2) analytic and synthetic algorithms, more or less independent of the languages chosen for input and output; (3) a descriptive grammar of Russian physics; (4) a Russian-English physics glossary. Now we have dropped (3) and (4), despite their importance, because the grammar and glossary that we have are incomplete. The algorithm is only a mirror, reflecting a descriptive grammar; the main products of MT research belong to linguistics, and a much smaller fraction to computer programming. Nevertheless, we have not processed enough text to justify our publishing a Russian grammar or dictionary. Several million running words of text are required as a suitable basis.

This paper is a sampling of our intended book. As readers of our reports already know, our method of research consists of processing text and analyzing the results.[3] We translate a new corpus mechanically, postedit it, and see what the posteditors have added. In this sampling I wish to emphasize the last two steps: postediting and analysis. We have been forced to take more pains with these steps than with the others because they have proved more difficult to organize to our satisfaction. Accordingly, we have been slow in reporting them.[4] As for substantive results, I will mention only the concept of idiom-in-structure, a concept that turns up in several places in the literature without ever getting quite the explication that it demands. Finally, I will introduce two topics that will soon occupy most of our time (as we expect), grammatic transformations and distributional semantics.

Postediting

After some trial and error, we reached the conclusion that postediting must include structural description of the input text. In the usual discussion, the editor is said to polish the rough

---

[3] Edmundson, H. P., and David G. Hays, "Research Methodology for Machine Translation." Mechanical Translation, vol. 5, no. 1 (July, 1958), pp. 8-15.

[4] Harper, Kenneth E. , David G. Hays, and Barbara J. Scott, Studies in Machine Translation - 8: Manual for Postediting Russian Text, Santa Monica, Calif. : The RAND Corporation Paper P-1624, Revised, 7 November 1959-

translation;[5] we do not gainsay that requirement, but we think it leaves an unnecessarily difficult job for the analyst who comes after the editor in the research process.

Our conception of syntactic structure has grown from month to month, and it is certainly not static at this time. We first thought of structure as a set of connections among the words in a sentence; Yngve, among others, urged us to go further, and Hiž's paper[6] at the Cleveland Standards Conference states, more clearly than we had ever put them, the reasons for more refined description. Roughly speaking, connection structure is an inadequate framework because it does not differentiate among dependents of the same governor. Yet two dependents of a single governor often have different functions; to reflect such differences, we turned to the grammar code.[7]

In our syntactic theory, both the subject and the object of a finite, transitive Russian verb depend on it. Their different functions are shown in their cases: one is nominative, the other accusative (or some other case, but not nominative). To each noun form in our glossary, we attach a grammar-code symbol that shows its case (or cases). The case of a noun does not specify its function, since an occurrence of an accusative noun can be, for example, the object of either a verb or preposition. The case and dependency situation of an occurrence, taken together, describe its function more precisely than either alone.

The case of a noun is often ambiguous morphologically. What if a nominative-or-accusative noun occurs as the dependent of a

---

[5] One of the most recent statements of the standard viewpoint (although the author's position coincides with ours, regarding the posteditor as an informant or data source) is: I. G. Mattingly, "Post-Editing for Feedback", Section I of Report No. NSF-3, Mathematical Linguistics and Automatic Translation (A. G. Oettinger, principal investigator), submitted by the Computation Laboratory, Harvard University to the National Science Foundation, August, 1959.

[6] Hiž, H. , "Steps toward Grammatical Recognition", presented at An International Conference for Standards on a Common Language for Machine Searching and Translation, Cleveland, Ohio, Sept. 6-12, 1959.

[7] Our codification of Russian grammar is described only in an obsolete report (now being revised): Kenneth E.Harper and David G. Hays, Studies in Machine Translation - 6: Manual for Coding Russian Inflectional Grammar, Santa Monica, Calif. : The RAND Corporation, March 3, 1958.

finite verb of which it can be either subject or object? Eventually we must have an algorithm to decide the function of each such occurrence;[8] meanwhile the posteditor can decide—giving us data from which to derive our algorithm.

The editor needs a notational scheme; hence we introduced the resultant grammar code, or RGC. The RGC has the same format, and virtually the same list of symbols, as the original, morphological grammar code (GC) used in the glossary. Both GC and RGC consist of five IBM characters; the third character, for a noun, shows its possible case-number uses. For example, the symbol "G" indicates genitive singular or nominative or accusative plural, while "5" means nominative plural, and "M" means accusative plural. By changing "G" to "5" or "M", an editor can indicate his decision about the function of a noun occurrence depending on a verb. Our routine for sentence-structure determination now modifies grammar-code symbols; the editor has only to correct the output of the machine, revealing shortcomings of our existing rules.

Even our current descriptive framework is incomplete. Every occurrence of a Russian genitive noun, depending on another noun, has the same function in our present notation. The Soviet Academy's "Grammar of the Russian Language" lists a number of distinct functions, subclassifying what we treat as a single function.[9] In English, the possessive has at least the three functions that Jesperson[10] pointed out, namely, possessive, subjective, and objective. Jesperson used transformations (without calling them by that name) to distinguish between subjective and objective: " 'England's wrongs' generally means the wrongs done to England".[11] Worth[12] explicitly relies on

[8] Cf. Hays, David G. , "Order of Subject and Object in Scientific Russian when other Differentia are Lacking", Mechanical Translation, in press.
[9] Academy of Sciences of the USSR, Grammatika Russkogo Yazyka, Vol. II, pp. 234-241.
[10] Jesperson, Otto, Growth and Structure of the English Language , Garden City, N.Y.: Doubleday (Anchor Books), p. 193.
[11] Loc. cit.
[12] Worth, Dean S. , "Transform Analysis of Russian Instrumental Constructions", Word, vol. 14, no. 2/3 (Aug. - Dec., 1958), pp. 247-290.

transformational possibilities to differentiate seven main classes of functions of Russian instrumental nouns, all depending on verbs or other nouns.

We are now at the point of looking for good analyses of functions lit this new level; almost ready to attempt our own analyses if we cannot find them in the literature; about to establish notational conventions for the posteditors to use, so that the posteditor can describe syntactic structures to this degree of precision. When the editors have processed enough text, we will be prepared to attempt automatic resolutions, but not before.

Semi-Automatic Analysis

Our reason for imposing such stringent demands on the editor ih that it makes analysis easier. For MT, an analyst must answer such questions as “When do we insert ‘of’ in the English translation **of** a Russian sentence?” His answer may be, for example, that “of” in inserted before the English equivalent of a Russian noun if the occurrence to the left is a noun, or before the English equivalent of a Russian adjective preceding a Russian genitive noun, if the occurrence to the left of the adjective is a noun, etc., always providing that the preceding noun is not translated by an English gerund, or by a noun in a certain lexical class, etc. We find it more convenient to factor this rule into several simpler parts. One section of the complex rule is devoted to finding the governor and function of the Russian noun; this section goes into our sentence - structure determination routine. The rest of it has to do with the exact English function of the genitive noun’s equivalent.

Factoring simplifies both the MT routine and analysis; here we are interested in the latter. After a corpus has been edited, the analyst orders certain listings. Concordances, or exhaustive listings of text by lexical and grammatic properties, are used, but we also use selective listings. Let us take specific examples.

First, consider the word что = “that” or “which”. We know from standard grammars[13] and our own experience that что has two distinct functions. This word is sometimes a subordinate conjunction (91% of its occurrences in our text) and sometimes a relative

[13] E.g.; Unbegaun, B. O., Russian Grammar, Oxford; Clarendon Press, 1957, pp. 128 and 277.

pronoun (9%). It is a conjunction only when it introduces a clause serving (i) as the object of a verb in a lexically limited class, (ii) as the object of an adjective or noun derived from such a verb, or (iii) in apposition with such a noun. The verbs, adjectives, and nouns belonging to this class are marked in our glossary (in the grammar code). That is to say, they are marked as soon as they are seen to govern что as a subordinate conjunction (with English equivalent "that"). It would be vain to try to mark them in advance; comparing the list we now have with Ushakov's dictionary,[14] Joan Pustula found that only half of the words on our list are marked as having the property in question.

When a word first occurs as governor of a noun clause introduced by что, it lacks the mark that would allow the sentence-structure routine to make a connection. A posteditor has to mark the occurrence of что as a dependent of the new word. The analyst calls for a list of occurrences of что as subordinate conjunction. The list is prepared in two sections: governor marked, governor unmarked. The first section is needed only for statistical records, the second because the governors listed these must be marked. Within each section, the list is ordered by the word number of the governor—i.e., alphabetically.

When the analyst receives the list, he reads through the new governors and checks context if he is surprised by what he finds. Checking is easy, because each entry in the list includes a reference to text location. Checking is needed because posteditors do make errors. After checking, the analyst prepares a list of glossary-change notices that go into an automatic updating procedure. The grammar-code symbols of the words just discovered to govern что as a subordinate conjunction are changed in the glossary; the next time one of them occurs in this combination, the sentence-structure determination routine will make the connection, if all goes well.

It is easy to make a process like this wholly automatic, especially if a suitable programming language is available. The

[14] Ushakov, D. N. , Tolkovyj Slovar' Russkogo Yazyka, Moscow 1935-1940 (4 vols.).

analyst can write a routine to search the text for relevant occurrences, inspecting the posteditor's marks as well as the product of the MT operation. He would have the routine print a list of new что - governors for verification and compile whatever statistics he wanted. If he dares to omit verification, the whole process, from postediting to glossary updating, can be automatic.

Identifying new governors of что = "that" is an example of adding new items to known categories. New categories remain to be discovered; automation of that process is more challenging.

Automatic operations can generate listings of exceptions to established rules. For example, if the relative adverb где = "where" is assumed always to modify a clause or verb, so that its governor is always a clause head, a list of exceptions would, in time, bring to light a class of words like случаи = "case", nouns that Russian physicists modify with где - clauses. If a certain preposition with a governor belonging to class X, say, and dependent (noun object) in case C and lexical class Y, is supposed to be translated with a certain English equivalent, and is given a different equivalent occasionally, a list of exceptions could eventually bring to light the existence of a subclass of X or a subclass of Y.

Automatic routines can also test for the existence of classes with certain defining properties (somewhat in the spirit of Giuliano's formula finder[15]). Can Russian adverbs be divided into two classes: those that always precede and those that always are preceded by their governors? With a convenient programming language—and using post-edited text with a syntactic description built in—it will be easy to answer such questions.

Many authors have higher goals than these. Andreyev[16] and Solomonoff [17] are only two of several who want programs capable of

[15] Giuliano, Vincent E. , and Oettinger, Anthony G., "Research on Auto matic Translation at the Harvard Computation Laboratory", presented at the International Conference on Information Processing, Paris, June, 1959.

[16] Andreyev, N. D., and Fitialov, S. Ya. , "Intermediary Language for Machine Translation and Principles of its Construction", in Abstracts of the Conference on Mathematical Linguistics, Leningrad, 15-21 April 1959, translated by U. S. Joint Publications Research Service, JPRS: 893-D.

[17] Solomonoff, R. J., The Mechanization of Linguistic Learning, Cambridge, Mass., Zator Co. , Report No. ZTB-125, April, 1959.

establishing translation algorithms with no more than parallel texts for input. Although such a program will probably be designed eventually, we believe that we are pursuing a more productive course for the present, and a more efficient policy for "known" languages.

Our method, in summary, consists of textual analysis. The posteditors serve as informants in a restricted sense; they supply "correct" sentence-structure analyses and "correct" translations, to be used as objectives. The analysts generate hypotheses, define new categories, and postulate new rules. The computing staff relieves us of time-consuming searches through text, makes statistical tabulations, keeps the glossary up-to-date, and generally eases the linguist's work. When Kelly's MIMIC system is far enough advanced, we expect to move rapidly toward fully automatic analysis within our present framework.

The Idiom-in-Structure

An idiom is a phrase, or sequence of forms, that comprises a lexical unit. The forms and their order are fixed; if the same forms occur in a different order, or with other occurrences intervening, the idiom is not recognized. Hundreds of idioms are listed in our glossary, but there are other frozen word combinations that we cannot call "idioms" because the occurrence order of their elements is variable.

Some of these semi-idioms consist of governor and preposition: зависимостъ от = "depend on", состоятъ из = "consist of", etc. Kenneth E. Harper, using the RAND corpora, has made a detailed study of this phenomenon in Russian physics; and the same idea turns up in standard grammars[18] (where the term "lexically closed", or "limited" is applied), in Russian work on MT,[19] and in Bar-Hillel's discussion[20] of discontinuous constituents in English (e.g. , "give up

[18] E.g., Academy of Sciences of the USSR, Grammatika Russkogo Yazyka, Vol. II, pp. 173-176.

[19] Belokrinitskaya, S. S., "Principles in Compiling a German-Russian Glossary of Polysemants for Machine Translation", in Abstracts of the Conference on Machine Translation, May 15-21, 1958, translated by U. S. Joint Publications Research Service JPRS/DC-241.

[20] Bar-Hillel, Yehoshua, Report on the State of Machine Translation in the United States and Great Britain, Jerusalem, Israel: Hebrew University, Feb. 15, 1959.

candy" or "give it up"). Other semi-idioms consist of verb and direct object, or verb and modifying phrase: играть роль = "play a role", иметь в виду = "have in view".

The frozenness of these expressions is often important for one reason or another. Selection of the English equivalent is frequently determined for one member of the combination by the other; Harper's objective has been to produce accurate translations of Russian prepositions. Syntactic functions can be influenced by the fact of combination; иметь governs что - clauses only when it is combined with в виду. Sometimes the fixity of the combination clarifies a structural ambiguity; we found that the subject and object of a verb, when they cannot be differentiated morphologically, can be distinguished by word order except when one of them is frozen in combination with the verb.[21] Thus we found имеет место правило = "a rule occurs" in our text, but nowhere did we find the order verb-object-subject when object and subject were morphologically identical and neither was closely associated with the verb.

Idiom-recognition routines, upon finding an occurrence in text that could be the first element of an idiom, look at the next following occurrence to see if it continues the same idiom, and so proceed. If an occurrence intervenes that does not belong to the idiom, recognition fails. Sentence-structure-determination routines necessarily span longer sections of each sentence. When sentence-structure determination is completed successfully, the components of the semi-idioms described above stand in frozen relation to each other; their sequence is free, but their structural connections are fixed. We propose, therefore, to establish a list of idioms-in-structure. The routine to recognize them would operate after sentence structure had been determined, and it would follow structural connections. On finding an occurrence that can be the dominant element of an idiom-in-structure, the routine would test the dependent occurrences to see whether one of them continued the idiom.

As we have shown, others discuss these semi-idioms, but we feel that our analogy with the ordinary idiom routine will be more efficient than the alternatives that are now in use.

---

[21] Op. cit., fn. 8

Grammatic Transformations and Distributional Semantics

The concept of transformations is now familiar, but there are no exhaustive lists of the transformations used in any natural language. Making these lists is a matter for empirical research; as yet no operational definition has been suggested that is entirely satisfactory as a basis for semi-automatic data analysis. Probably Harris came closest in his original paper on the subject.[22] Confronted with the high cost of language-data processing, however, he immediately dismissed his own suggestion. The cost of automatic processing is rapidly decreasing, and it will decrease much faster when automatic print readers are generally available. As our systems for sentence-structure determination (algorithms, dictionaries, and grammars) are perfected, so that posteditors have less to add, we can foresee empirical studies based on tens of millions of running words of text. Anticipating that time, we propose to undertake preliminary empirical searches for transformations, using the Russian physics text that we have or can acquire.

In this study, we will test an operational definition of transformation that is appropriate to our dependency theory of syntax. Abstractly, a transformation is defined as a pair of dependency types, linking different grammatic types, but equivalent in meaning. Consider four grammatic classes: W, X, Y, and Z. Suppose that in our text occurrences of class X govern occurrences of class W (WdX), and that occurrences of class Z govern occurrences of class Y (YdZ). Because the grammatic classes involved are different, we can speak of two types of connections: Wd_1X and Yd_2Z . For example, nominative nouns depend as subjects on verbs, and adjectives depend as modifiers on nouns. We will say that d_1 and d_2 are coupled by a transformation if, for every $w_i d_1 x_j$ in our text (every pair of words of classes W and X that are connected, somewhere in our text, by d_1), there exists in our text a $y_i d_2 Z_j$ such that $w_i = y_i$, and $x_j = z_j$. In other words, “He wronged England” and “England’s wrongs” are, together, evidence that Nd_1V and $N_{poss}d_2N$ are transformationally equivalent. If all the evidence of a corpus supports equivalence, we believe it.

[22] Harris, Zellig S., “Discourse Analysis”, Language, vol. 28, no. 1 (1952), pp. 1-30.

The first and obvious difficulty is that a finite corpus demands a statistical measure. If d_1 and d_2 are transformationally equivalent, then in a large corpus many (but not all) of the word pairs connected by d_1 should also appear connected by d_2 . We need measures to tell us whether the observations that we make support the hypothesis of equivalence to an adequate degree.

The second difficulty, equally obvious, is that most of the intuitively equivalent pairs of connections bind pairs of words of different word classes. In "He wronged England" and "England's wrongs" we have $N_{obj}d_1V = N_{poss}d_2N$. Derivational families must be established so that $w_i = y_i$, can be interpreted as "w_i and y_i, belong to the same derivational family". The RAND glossary groups forms into words; now it is necessary to group words into derivational families. Dean S. Worth of UCLA is interested in this problem, and is planning a study based in part on the RAND corpora. We hope to incorporate his results into our study of transformations.[23]

Partial equivalence raises a third difficulty. As Jesperson noted, the English possessive is both subjective and objective; in Russian, too, both $N_{nom}dV$ and $N_{acc}dV$ are transformed into $N_{gen}dN$. Verbs and nouns have to be classified, if a given occurrence of $N_{gen}dN$ is to be identified as subjective or objective. Let N^x be a class of nouns that can serve as subjects of verbs in class V^y, and let N^x be a class of nouns that can serve as objects of the same verbs. Let N^y be a class of nouns derived from verbs in V^y. Then $N^x{}_{gen}dN^y$ is a transformation of $N^x{}_{nom}dV^y$, and $N^z{}_{gen}dN^y$ is a transformation of $N^x{}_{acc}dV^y$. If we begin by looking at $N_{nom}dV$ as a homogeneous class, we should find that every pair of this type corresponds to a pair $N_{gen}dN$, but not vice versa. Therefore we should be led to distinguish subjective genitives, and to examine the others separately.

The difficulties go on without end for as far as we can now see. We anticipate a long and interesting task in the development of empirical methods for research on transformation.

---

[23] The use of transformations in establishing word families is proposed by Z. M. Volotskaya, "Word Formation in Conversion of Intermediary Language into Output Language", in op. cit. , fn. 16.

Our other task for the immediate future is a study of distributional semantics. Harris put it this way: "If the environments of A are always different in some regular way from the environments of B, we state some relation between A and B depending on this regular type of difference ... If A and B have almost identical environments except chiefly for sentences which contain both, we say they are synonyms ... If A and B have some environments in common and some not, . . . we say that they have different meanings, the amount of meaning difference corresponding roughly to the amount of difference in their environments"[24]. We hope to follow this program as far as it can take us, changing it a little to adapt it to a dependency theory of sentence structure.

Harper is getting started with a study of verbs, separating those with only animate subjects in our text, those with only inanimate subjects, and those with both. That step is only a beginning, and where it will lead is unknown. One thing seems certain: we will use distributional semantics to establish word classes, apply those word classes in studying transformation, and use transformation analysis in the study of semantic distribution.

Conclusions

We are writing a major report of our results to date. We are anxious to promote automatic programming for the sake of easier analyses of the material we are collecting. And we are fretting under the MT label.

The report will show how we translate Russian physics text into English, and it will contain both samples of the output and measures of our efficiency and effectiveness. The system that we use can be improved, and we hope that we and others will improve it. The code-matching method of Parker-Rhodes, Lukjanow, Garvin,[25] et. al. , would improve our system considerably, and either the Lamb-

[24] Harris, Zellig S. , "Distributional Structure", Word, vol. 10 (1954), pp. 146-162. The quoted passage is on p. 157.

[25] Parker-Rhodes, A.F., "An Algebraic Thesaurus", presented at an International Conference on Mechanical Translation, Cambridge, Mass., Oct. 15-20, 1956. Ariadne Lukjanow and Paul Garvin have (independently) communicated their interest in code-matching techniques, in private conversations.

Jacobsen[26] or the Ziehe-Kelly[27] glossary-lookup method has to be built in. The most important means of improvement, of course, is enlargement of the textual base of the dictionary and grammar.

Automatic programming is often regarded as a panacea and used to cure problems that could better be attacked by explication. There is a limit to our powers of explication, however, and MIMIC is a token of Kelly's success in advancing automatic programming in the RAND MT project. Its first use, as he explains in this Symposium,[28] is for output construction. We also plan to use it in Netting up data-reduction operations, and eventually in programming transformational manipulations of our text. To serve these purposes, MIMIC must grow.

Until late 1959 we accepted the label "MT", but two months ago we petitioned for a change. Our new titles are linguistic research and automatic language-data processing. These phrases cover MT, but they allow scope for other applications and for basic research.

Machine translation is no doubt the easiest form of automatic language-data processing, but it is probably one of the least important. We are taking the first steps toward a revolutionary change in methods of handling every kind of natural-language material. The several branches of applied linguistics have so much in common that their mutual self-isolation would be disastrous. The name of our journal, the name of our society if one is established, the scope of our invitation lists when we meet, and all other definitions of our field should be broadened—never narrowed. In 10 years we will find that MT is too routine to be interesting to ourselves or to others. Applied linguistic research is endless.

---

[26] Sydney M. Lamb and William H. Jacobsen, Jr. , personal communications.

[27] Kelly, Hugh S., and Theodore W. Ziehe, "Glossary Lookup Made Easy", in this Symposium.

[28] Kelly, Hugh S., "MIMIC: A Translator for English Coding", in this Symposium.